Nomage: an electronic lexicon of French deverbal nouns based on a semantically annotated corpus

A Balvet, L Barque, M.-H Condette, Pauline Haas, R Huyghe, R Marín, A Merlo

Nomage: an electronic lexicon of French deverbal nouns based on a semantically annotated corpus

A. Balvet*, L. Barque, M.-H. Condette*, P. Haas*, R. Huyghe***, R. Marín*, A. Merlo***

*STL, CNRS UMR 8163, **LDI, CNRS UMR 7187, ***EILA
U. Lille 3, U. Paris 13, U. Paris 7
prenom.nom@univ-lille3.fr, pnom@ldi.univ-paris13.fr, pnom@eila.univ-paris-diderot.fr

Abstract

ANR-funded Nomage project aims at describing the aspectual properties of deverbal nouns taken from a corpus, in an empirical way. It is centered on the development of two resources: a semantically and syntactically annotated corpus of deverbal nouns based on the French Treebank, and an electronic lexicon, providing descriptions of morphological, syntactic and semantic properties of the deverbal nouns found in our corpus. Both resources are presented in this paper, with a focus on the comparison between corpus data and lexicon data.

1. Introduction

From a theoretical standpoint, the works of (Lees, 1960), through (Chomsky, 1970) and (Grimshaw, 1990), provide a laying ground for our description of deverbal nouns' properties, though these works focus mainly on morphological and syntactic aspects. We elaborate on this theoretical framework, by providing fine-grained descriptions of the morphological, syntactic, semantic (more precisely aspectual) properties of deverbal nouns in an empirical way. In this paper, after a brief revision of related work (section 2.), we present the Nomage corpus and the semantic annotation process applied to deverbal nouns (section 3.). We then present the structure and content of our lexicon, which describes the deverbal nouns extracted from our corpus, alongside the morphologically-related verbs we manually associated to each of these nouns (section 4.). Since, in our project, the description of deverbal nouns is carried out by means of two different methods, in the last section we confront annotations taken from the corpus with those taken from the lexicon (section 5.).

2. Related work

Leaving aside Verbaction (Tanguy and Hathout, 2002), an xml database of nominalizations paired with their verbal bases, the resource we present here is, as far as we know, the first attempt to semantically annotate both a corpus and a lexicon of French deverbal nouns.

Similar resources exist for other languages, particularly for English and Spanish. For English, the most relevant resource is NOMLEX, a lexicon of English deverbal nominalizations containing 1,025 entries (Macleod et al., 1998). It is mainly focused on argument structure: the allowed complements of nominalizations are described and linked to their corresponding verbal arguments. NOMLEX-PLUS (Meyers et al., 2004), an integral part of the NomBank project (Meyers, 2007), is an extension of NOMLEX. It includes 7,050 additional entries: 4,900 for verbs' nominalizations, 550 for adjectives' nominalizations, and 1,600 corresponding to other argument-taking nouns.

For Spanish, one can cite AnCora-Nom (Peris et al., 2010), a lexicon of 1,655 lexical entries corresponding to the different deverbal nominalizations appearing in the annotated corpus Ancora-Es (Taulé et al., 2008). Ancora-Nom not only includes information on argument structure, like NOMLEX, but also on lexical aspect.

3. The corpus

In this section, we outline the main features of the electronic corpus we use, the French Treebank, and we describe the deverbal noun candidates' extraction process. Then, we proceed by describing our semantic annotation protocol.

The French Treebank is a 1 million words corpus of newspaper articles taken from *Le Monde*. It provides several levels of linguistic annotations: simple and compound tokenization, lemmas, part-of-speech tags augmented with morphological information, together with constituent boundaries and syntactic functions for half of the corpus (Abeillé et al., 2003).

Based on morphological cues (suffixes: -ion, -age, -ment, etc.), we extracted over 10,000 nominalization candidates (simple tokens only) from the functionnaly-annotated half of the French Treebank. After close inspection, only 4,042 candidates were considered in the course of the project: all nouns that were not syntactic heads (e.g. *un permis de* ***construction*** *(a construction permit)* versus *la* ***construction*** *européenne (the European integration)*) of a NP were discarded, because of their incompatibility with the transformation tests we used for the semantic annotation process (see below, section 3.2.). Moreover, some nominalizations stem from an adjectival base, and not a verbal one: e.g. INDULGENCE stems from INDULGENT, but our project aims exclusively at deverbal nouns. The Nomage project is dedicated precisely to the study of the inheritance of semantic and aspectual features from the verbal bases, thus some of the extracted candidates, which were possible converted nouns, were also discarded. Even though a link to a verbal lexeme can be found, the directionality of the inheritance relationship cannot be clearly established; this includes cases such as VOYAGE (*travel*, noun) and VOYAGER (*travel*, verb). Finally, some amount of noise is attributable to the extraction process itself. Morphological cues in themselves do not discriminate between true nominaliza-

tions and false-positives: items such as SARCOPHAGE (*sarcophagus*) had to be filtered-out, based either on an automatic filtering ("stop-list" lookup) or a manual process.

3.1. Aspectual annotation of deverbal nouns

One of the central methodological features of project Nomage is that the semantic descriptions rely on the application of transformation tests, carried out by "naive" annotators and not on forged examples. These tests were devised so as to highlight a selection of semantic properties for each candidate: its aspectual structure, together with its mass/count status. We wish to emphasize here that the transformational tests were intentionally devised so as to be applied by native speakers that had received no training in linguistics. These annotators were not aware of the fine-grained semantic and aspectual distinctions we were trying to describe, but rather they were simply asked to assess whether each transformation yielded acceptable sentences or not.

Originally, we had planned to implement a cross-annotation process for each candidate, in order to provide minimal inter-annotator agreement (kappa) assessment. Unfortunately, due to a lack of available annotators, this methodology had to be abandoned: up to 7 "naive" annotators were hired for this project, some of them at different points in time, working on partially intersecting annotation batches, while a minimum of 10 *distinct* annotators would have been required. Moreover, due to data integrity issues, part of the candidates had to be manually corrected by researchers (and thus far from "naive") associated to the project. Therefore, it is not possible to provide inter-annotator agreement scores for our data.

3.2. Using transformational tests to assess semantic properties

The transformational tests were voluntarily presented in an unstructured manner to the annotators, so as to avoid any implicit theory-forming on their part. We present below the semantic annotation of nominalization ÉVALUATION, based on our methodology.

*L'**évaluation** faite selon les critères du BIT (Bureau International du travail) n'est pas plus rassurante. [The evaluation carried out by the BIT is not more reassuring.]*
T1_Plusieurs : yes : → *Plusieurs évaluations* **T2_Avoir lieu :** yes → *L'évaluation qui a eu lieu hier* **T3_Éprouver/ressentir :** no **T4_Un peu de :** no **T5_Durer x temps :** yes → *L'évaluation qui a duré 2 jours* **T6_Se trouver (qq part) :** yes → *L'évaluation qui se trouve sur ton bureau* **T7_Effectuer/procéder :** yes → *L'évaluation effectuée hier* **T8_État de :** no **T9_Se dérouler :** yes → *L'évaluation qui s'est déroulée hier* **T10_Card :** yes → *Deux évaluations*

Table 1: Semantic annotation of *évaluation*

Our tests allow us to uncover two main semantic features: mass/count status, and aspectual structure. Annotators had to assess whether the original determiner could be replaced by *plusieurs* 'several' (test 1), *un peu de* 'some' (test 4) or by a cardinal determiner (test 10). Here, tests 1 and 10 yield a positive outcome, while test 4 is impossible, which allows us to categorize ÉVALUATION as a count noun.

As for aspectual properties, *avoir lieu* 'happen/hold' (test 2) and *effectuer/procéder* 'complete/perform' (test 7) are meant to identify whether the candidate has an **event** reading. Here, it is precisely the case: both tests can be applied. In addition, "se dérouler" (test 9) indicates that the considered noun is a durative event. Other tests are aimed at non-event readings: tests such as *éprouver/ressentir* 'feel' (test 3) and *état de* 'to be in a state of' (test 8)[1] allow us to identify **state** readings. Here, this occurrence of ÉVALUATION is not compatible with these latter tests, which is, in itself, a confirmation of its event reading. Finally, *se trouver (qq part)* 'to stand/be located at' is meant for capturing **object** readings.

3.3. Test outcomes and semantic categorization

Test outcomes on our 4,042 items are interpreted so as to yield 3 classes: EVT (events), ETAT (states) and OBJET (objects). In order to be categorized as a state, a candidate must exhibit at least one positive outcome for tests 3 or 8. For objects, only test 6 is considered, while for events a candidate must yield one positive outcome for test 2. Therefore, even though our tests may appear partly redundant, this is intentional, as some tests are considered as more generic and others more specific. In the case of events for instance, test 2 is more generic than test 9, it is thus more discriminating: "avoir lieu" allows us to distinguish event and non-event readings, while test 9 allows us to further specify an event subclass. Moreover, this design serves as a rough control mechanism so as to avoid inconsistencies in annotations: for example a positive outcome to test 9 is supposed to entail a positive outcome for test 2. Annotations that do not follow this pattern are easy to spot and are put under close scrutiny in the final validation process. As for test 5, it is used along with test 9 to discriminate a certain subclass of events –the durative ones as opposed to the punctual ones. But test 5 is also valid for states[2] and in some cases may help categorize them. As can be seen in table 2, the conjunction of different test outcomes is used to yield "inferred" semantic classes, which will be compared to hand-coded semantic classes in the lexicon, in section 5. Examples (1a) through (1c) and table 2 give an illustration of the semantic classes that can be associated to each occurrence, based on their respective test outcomes, as coded by our naive annotators.

(1) a. *L'**évaluation** faite selon les critères du BIT (Bureau International du travail) n'est pas plus ras-*

[1] For this test, the sequence "état de" has to be inserted between the candidate and its determiner : * *L'**état d**'évaluation faite selon le BIT* ...

[2] Test 5, to some extent, is also valid for objects (e.g. *Sa télé a duré 2 mois avant de tomber en panne*) but has not the same interpretation.

surante. [The evaluation carried out by the BIT is not more reassuring.]

b. *Il s'agit de produits récupérés à l'état liquide dans les **installations** de traitements des gaz. [This refers to liquid-state products recovered from gas processing facilities.]*

c. *Le **mécontentement** est de plus en plus grand en Pologne à la suite des fortes hausses des prix du gaz, de l'electricité et de l'eau chaude appliquées au début de l'année. [Discontent grows in Poland following a sharp increase in gas, electricity and hot water prices.]*

	(1a)	(1b)	(1c)
2. *Avoir lieu*	**yes**	no	no
3. *Éprouver*	no	no	**yes**
5. *Durer x temps*	**yes**	no	**yes**
6. *Se trouver*	**yes**	**yes**	no
7. *Effectuer/procéder*	**yes**	no	no
8. *État de*	no	no	no
9. *Se dérouler*	**yes**	no	no
Inferred class	EVT or OBJET	OBJECT	STATE

Table 2: Interpretation of aspectual test outcomes

As can be seen, based on test outcomes (table 2), the occurrence of EVALUATION in (1a) has two related meanings, an action and its result, that can be co-predicated in the same sentence (Pustejovsky, 1995; Godard and Jayez, 1996; Milicévic and Polguère, 2010).

4. The lexicon

4.1. A lexicon entry

The Nomage lexicon describes each deverbal noun from our corpus (amounting to 746 nominal lexemes)[3], as well as their verbal base (679 verbal lexemes). Each nominal lexeme is associated with an aspectual class and a semantic argument structure. Note that the aspectual class is not attributed to lexemes according to the results of the tests applied to their occurrences in the corpus (see section 3. above) but following a classical method that will be explained in section 4.1.2. below. We emphasize here that our goal is precisely to contrast two aspectual annotation methodologies.

Tables 3, 4 and 5 below illustrate the kind of information that can be found in the Nomage lexicon, with the description of AMÉNAGEMENT#1 and its verbal counterpart AMÉNAGER#1. As illustrated in table 3, AMÉNAGEMENT#1 has two arguments (X and Y) and denotes an accomplishment (i.e. a durative event).[4]

[3]These 746 nominal lexemes correspond to the 4,042 tokens in the corpus. The average number of examples per lexemes is thus 5.5.

[4]The aspectual classes assigned to each lexeme are based on a finer-grained ontology than the habitual three classes (EVT, STATE, OBJECT). In our lexicon, we distinguish for instance durative events from non durative ones, and telic from non telic ones (see below section 4.1.2.).

id	45
Lexeme	AMÉNAGEMENT#1
Argument structure	$\sim$ de Y par X
Aspectual class	ACC
Occurrences in the FT	{id:1794 ; id:1929}
Verbal base	id:44

Table 3: Description of noun AMÉNAGEMENT#1

Alongside the information given above, each entry points to a verbal source. It is thus possible to have access to a description of the verbal lexeme AMÉNAGER#1 through the nominal one AMÉNAGEMENT#1. As can be seen in table 4, the verb's argument structure and aspectual class are also described in our lexicon.

id	44
Lexeme	AMÉNAGER#1
Argument structure	X $\sim$ Y
Aspectual class	ACC

Table 4: Description of verb AMÉNAGER#1

Finally, each entry is associated with its corresponding occurrences in the original corpus, and the actual realization of the lexeme's arguments (table 5).

id	1794
Deverbal	id:45
Occ.	Tout ce travail préparatoire sera fondamental pour l'**aménagement** universitaire au cours des cinq prochaines années.
Réal. Arg.	X:$\emptyset$, Y:adj. rel.

Table 5: An occurrence of AMÉNAGEMENT#1

4.1.1. Argument structure

In our lexicon, we describe the semantic arguments of each nominal and verbal predicates in a systematic manner. By semantic arguments we mean the required participants in order to define the state of affairs denoted by the considered predicate (Mel'čuk, 2004a). Semantic arguments are represented by variables (X, Y, Z), as can be seen in the description of AMÉNAGEMENT#1, which is associated with two arguments X and Y. The Dicovalence lexicon (Van den Eynde and Mertens, 2003) frequently helped us to identify the semantic arguments of verbal predicates, which are generally also those of the corresponding nominal predicate. This is the case for AMÉNAGER/AMÉNAGEMENT: X represents in both cases the "agent" and Y the "undergoer". Each lexeme is associated with a description of the surface realization of its semantic arguments in the corpus[5] (Mel'čuk, 2004b). Lexeme AMÉNAGEMENT#1 occurs for example in the following sentences of the corpus :

[5]Note that not all possible realizations of a given semantic argument structure are described: we only consider the realizations found in our corpus.

(2) a. *Tout ce travail préparatoire sera fondamental pour l'**aménagement** universitaire au cours des cinq prochaines années.* X:∅, Y:adj. rel. (cf. table 5 above)

b. IBM devient ainsi actionnaire de Dassault systèmes à hauteur de 10% et assure la **commercialisation** de ses logiciels Catia. X:∅, Y:adj. rel., Verbe Support= X assurer det N

In sentence (2a), argument X of AMÉNAGEMENT#1 is not realized, while argument Y is realized by a relational adjective (*universitaire*). Note that arguments that are syntactically dependent from the light verb of a nominalization are also described: for example, semantic argument X of COMMERCIALISATION is realized as the subject of the light verb *assurer* in sentence 2a.

4.1.2. Aspectual class

We follow a classical approach to the description of the aspectual class of the deverbal nouns in our lexicon. We use aspectual tests, taken from the literature, in order to characterize their semantic and aspectual properties. In contrast, as has been shown above, the attribution of an aspectual class to each occurrence taken from our corpus was based on a set of transformational tests meant to be applied by "naive" annotators in the original context. We give a comparison between these annotation methods in section 5.

The first four labels retained are taken from Vendler's classification of verbs (1967), with slight adaptations, particularly by using the feature [+/- culminating], and extended to the nominal field. Lexemes of the states class (ETAT) denote non dynamic situations (e.g. POSSÉDER, ADMIRATION, etc.). On another branch of the aspectual ontology, lexemes of the activities class (ACT), such as MANIFESTER and PROMENADE denote dynamic, durative and non culminating situations. Accomplishments (ACC), such as RÉPARER and DÉMÉNAGEMENT, describe dynamic, durative and culminating situations. Finally, lexemes of achievement type (ACH) denote dynamic and culminating but non durative situations (e.g. ADOPTER and ACQUISITION).

The aspectual descriptions in our lexicon rely on original classes, as we have frequently observed that some lexemes do not match any of the simple classes mentioned above, but seem rather to constitute intermediate categories: thus, between achievements and states, we have proposed "stative achievements" (ACH-ETAT) which react positively to some tests dedicated to achievements but also to some tests accepted for states – particularly tests of duration, when these tests concern a resultant state. This class is dedicated to items such as EMPRISONNEMENT which denote an achievement (the sending to prison) followed by a state that lasts until the end of the process (the coming out of prison). In the same way, we propose "stative accomplishments" (ACC-ETAT) which describe an accomplishment followed by a state. This class encompasses cases such as INVASION which refers to the durative action of the invasion of a territory and to the state of occupation of the invaded territory. We have also introduced "accomplishments-activities" (ACC-ACT), which constitute an intermediate class between the ACT and the ACC, and denote activities of which each step could be considered as the final stage. This class comprises items such as REFROIDIR, RÉTRÉCISSEMENT, etc. This category is also known under the noun of "degree-achievement" (Dowty, 1979). The classes we have just presented apply at the same time to verbal and nominal lexemes. However, the existence of semantic idiosyncrasies in the nominal field has made us consider several new aspectual categories so as to label our nominalizations more finely.

More precisely, for the class of activities, we've had to add a label in the nominal field so as to take into account the fact that verbs of activity (e.g. JARDINER, SE PROMENER, MANIFESTER) do not yield a homogeneous class of nominalizations (Flaux and Van de Velde, 2000; Haas et al., 2008; Heyd and Knittel, 2009). Indeed, the opposition massive / countable distinguishes, at the aspectual level, two types of nouns derived from verbs of activity: countable nominalizations (e.g. PROMENADE) and massive nominalizations (e. g. JARDINAGE). From the aspectual point of view, all these nouns describe dynamic, durative and non culminating situations, but only count nouns denote actions which are temporally delimited, i.e. events (Haas and Huyghe, 2010). We keep the ACT label for these deverbal activity count nouns, which are statistically the most representative of the category, whereas their massive counterparts are labeled HAB (for "habitude"), because they can denote routine activities (Barque et al., 2009).

Another particularity of nouns is that, contrary to verbs, they can denote objects, and in this case they are devoid of any aspectual features. This property is known for the nominalizations that express the result of an action (Grimshaw, 1990), but it can be extended to a wider set of nominalizations. So we consider the existence of a class called OBJET, in which we group together nouns that denote material objects (e. g. CONSTRUCTION), nouns that denote objects with an informational content (e.g. AFFIRMATION), and nouns that denote entities which induce a psychological state (e.g. OBSESSION). Finally, we have used complex classes that include nominal lexemes which are likely to denote a situation and/or an object (Pustejovsky, 1995; Godard and Jayez, 1996; Milicévic and Polguère, 2010). These lexemes can receive co-predication, as in *Son exposé fut long et ennuyeux*, where *long*, which qualifies the presentation course and progress, applies to the "accomplishment" aspect of EXPOSÉ, whereas *ennuyeux*, which qualifies the informational content of the presentation, applies to the OBJET meaning. Such a case receives the ACC•OBJET label.

The tests for assigning an aspectual class to verbs are well known in the literature. But the aspectual properties of nouns have been less studied, so we've had to adapt the classical verb-oriented tests to this class of lexical units. The set of these tests, which are presented in detail in the documentation of the lexicon (written in French), is available at the following address: `http://nomage.recherche.univ-lille3.fr/` (attached in the "délivrables" part of the site).

Table 6 summarizes the different aspectual classes at-

tributed to each entry (nominal or verbal) in the lexicon.

Verbal classes	ACC, ACC-ETAT, ACH, ACH-ETAT, ACT, ACT-ACC, ETAT
Nominal classes	verbal classes + ACC•OBJET, ACH•OBJET, HAB, OBJET

Table 6: Aspectual classes in the lexicon

Table 11 shows the correspondance between the fine-grained classification used in the lexicon and the more general classification used in the corpus.

Corpus aspectual classes	lexicon aspectual classes
EVT	ACC, ACT, ACT-ACC, ACH, (ACC/ACH)-ETAT, (ACC/ACH)•OBJET
ETAT	ETAT, ACC-ETAT, ACH-ETAT
OBJET	OBJET, ACH•OBJET, ACC•OBJ

Table 7: Two sets of aspectual classes

4.2. Polysemy ratio

Table 8 shows the overall polysemy ratio of nominal and verbal forms from our lexicon.

Nominal lexemes	746
Nominal forms	656
Nominal polysemy ratio	1.14
Verbal lexemes	679
Verbal forms	648
Verbal polysemy ratio	1.04

Table 8: Polysemy ratio in the Nomage lexicon

The low polysemy ratio (1.14 lexemes by entry) can be explained by the fact that our corpus is relatively small (500,000 words) and specialized (newspaper articles). For deverbal nouns, polysemy comes from two main sources: it can either be inherited from the verbal base, or it can be attributed to the noun itself. For example, in the sentences below, each PROMOTION lexeme derives from two distinct PROMOUVOIR verbal lexemes.

(3) a. *C'est arrivé après sa **promotion** au poste de directeur financier.* (la personne X PROMOUVOIR#1 l'individu Y au poste Z → PROMOTION#1 de Y à Z accordée par X)

b. *Chirac va faire la **promotion** de son livre en plein marasme judiciaire.* (la personne X PROMOUVOIR#2 Y → PROMOTION#2 de Y par X)

In our lexicon, the other source of polysemy is mostly attributable to metonymy links that can be observed between an action and one of its participants or between an action and its result (Bisetto and Melloni, 2008). For instance, in our lexicon we describe two lexemes INSTALLATION, one denoting the fact of installing something, the other the result of the process (the installed thing).

5. Analysing data

5.1. Suffixes and aspectual classes in the lexicon

From a morphological point of view, one of the main descriptive and theoretical issues is the relationship between the aspectual class of a nominal lexeme and its suffix. The table below is a census of the different semantic class[6] → suffix mappings in our lexicon.

	EVT	STATE	OBJ	HAB	total
-ade	6	-	-	-	6
-age	45	2	7	2	56
-ance/-ence	10	19	8	2	39
-ée	13	2	3	-	18
-ion	336	36	61	12	445
-ment	133	12	14	3	162
-ure	11	1	6	2	20
total	554	72	99	21	746

Table 9: Distribution of aspectual classes by suffix

As can be seen, the most productive suffix is -ion (60.5%), followed by -ment (20.7%) and -age (7.6%). These results conform to those given by Tanguy & Hathout (2002). Regarding aspectual classes, events are the most frequent (75.3%) class, followed by objects (13.5%) and states (9.8%).
As for the relationship between suffix and aspectual class, we can notice that:

- -ance/-ence is the only suffix with less than 50% of events cases; this suffix also has the strongest tendency to combine with states. This result amends the rather widespread idea (Gaeta, 2002) that suffix -ance/-ence is only compatible with states. Our results show that, though it is true this suffix has a marked preference for states, it also combines with other aspectual classes (Dal and Namer, 2010).
- -age, -ée, -ment and -ion suffixes behave in similar fashions: between 70% and 80% of words bearing these suffixes are events.
- -ure offers fewer cases of events (55%); it is also the suffix which has the strongest tendency to combine with objects (30%).

Nevertheless, if we compute 95% confidence intervals with an error-rate of 0.05 based on figures of absolute frequencies over 100 occurrences, the size of the intervals is seldom under 7%. For instance, if we filter-out low-frequency suffix-aspectual class distributions and keep only those suffixes over 100 occurrences, the confidence interval for -ion as an EVT (336 occurrences) is [71.5;79.5]. For -ment, as an EVT (133 occurrences), it is [76.2;88][7]. Therefore, the

[6] As illustrated in table 11, we use two distinct set of aspectual classes : a fine-grained one to classify the lexemes and a more general one to classify occurrences of deverbal nouns in the corpus. The class HAB is the only one that can be generalized as EVENT, STATE or OBJECT.

[7] The 95% confidence intervals were computed based on a standard margin of error, following the function: Po $\pm$ 1.96 x

size of the confidence intervals is an indication that these figures are to be taken with extreme caution and should be computed on larger sets of data for higher confidence thresholds. Interestingly, χ^2 scores computed on these data show that the only suffix for which the null hypothesis should be discarded is -ance/-ence as a STATE[8]. Therefore, a strong connection between this suffix and the STATE class cannot be attributed to mere chance.

5.2. Verb → Noun inheritance of semantic properties

The main issue in this project is to assess whether a deverbal noun inherits (part of) the semantic and aspectual properties from the associated verbal form or not. In order to address this, we have assigned an aspectual class to each verb and noun described in our lexicon (see 4.1.2.), which enables us to compare and analyze matches and discrepancies between verbal and nominal domains. Our data indicate a perfect match between verbal and nominal aspectual class in around 67% of cases (492 perfect matches out of 737 verb-noun pairs). The remaining 245 verb-noun pairs exhibit at least some degree of discrepancy. Two main cases appear:

1. verbs and their nominalizations belong to two different classes entirely;
2. verbs and their nominalizations belong to slightly different classes.

5.2.1. Total verb-noun aspectual discrepancy

This case represents 73% (178 cases out of 245) of all mismatches, of which at least a partial explanation can be found in the existence of OBJECT classes for nouns, which by definition have no counterpart in the verbal domain. In this case, nominalizations do not denote an abstract situation (ACT, ACC, ETAT, etc.) but rather an object devoid of all aspectual properties. Around 55% of total discrepancies fall in this category (98 out of 178), for example: AGGLOMÉRER (ACC) → AGGLOMÉRATION (OBJET). The same holds for the HAB (routine activities) class for the nominal domain, which represents around 9% of the total discrepancy cases, e.g. RÉSISTER (ACT) → RÉSISTANCE (HAB). The remaining 64 verb-noun pairs (over 35%) are cases where the observed verb-noun aspectual mismatch cannot be explained by the existence of a class restricted to nouns: in some cases, only a slight discrepancy can be observed, e.g. INTERVENIR (ACC) → INTERVENTION (ACT) (in both cases we are dealing with durative events). In other cases, a major discrepancy can be observed, between the verbal and nominal domains, e.g. SOUFFRIR (ACT) → SOUFFRANCE (ETAT) (shift from dynamic to stative situation).

5.2.2. Partial verb-noun aspectual discrepancy

67 verb-noun pairs out of our 178 aspectual discrepancy cases are only partial mismatches. One of the causes for such mismatches is simply the overall discrepancy between verbal and nominal aspectual ontologies: as was presented above (4.1.2.), we propose complex aspectual classes such as ACH•OBJET, ACC•OBJET, etc., on the one hand, and complex classes such as ACH-ETAT and ACC-ETAT on the other hand. As for partial verb-noun aspectual discrepancies, we distinguish cases where:

1. the verb belongs to a complex aspectual class whereas the nominalization belongs to a simple class, which is a subclass of the verb's complex class. A "reduction" of the verb's complex aspectual class is thus at play; this is the case for over 37% of verb-noun pairs (25 out of 67), e.g. ACCUSER (ACH-ETAT) → ACCUSATION (ACH);
2. the noun belongs to a complex aspectual class where one of the subclasses corresponds to either a simple verbal class or one of the verbal complex class constituents. An elaboration on the verbal aspectual class is thus at play; this is the case for over 62% of verb-noun pairs (42 out of 67), e.g. DÉFINIR (ACC) → DÉFINITION (ACC•OBJET).

Verbal/nominal aspect correspondence		Total
Perfect match		492 (66.8%)
Mismatch	total	67 (9.1%)
	partial	64 (8.7%)
Other		114 (98 OBJ / 16 HAB)

Table 10: Verb-noun aspectual discrepancies

Cases summed up in the last line of table **??** are mismatches stemming from a difference between verbal and nominal aspectual ontology.

5.3. Comparing both methods of aspectual class attribution

In this project, we have used two different semantic annotation methods: one based on transformation tests applied on real-life sentences by naive annotators, the other based on forged sentences applied by linguistically trained annotators. In this section, we wish to assess whether both methods yield the same classes or not.

As can be seen in table 11, the degree of correspondence between aspectual classes assigned by each method is very high: for events, 2,001 matches out of 2,309 cases; for states, 136 matches out of 217, and for objects 211 out of 232.

CA	nb occ	distribution in lexicon
EVT	**2,309**	EVT (**2,001**), STATE (94), OBJECT (153), Other (61)
STATE	**217**	STATE (**136**), EVT (53), OBJECT (22), Other (6)
OBJECT	**232**	OBJECT (**211**), EVT (19), STATE (0), Other (2)

Table 11: Comparison of semantic class attribution based on two different methods

$\sqrt{((PoxQo)/n)}$, where Po is the percentage of the observed property, Qo the complementary percentage.

[8] Standard χ^2 with 18 degrees of freedom.

The differences stem in most cases from aspectual encoding errors in the lexicon. Thirteen occurrences of lexeme PROCÉDURE (in the sense of *legal procedure*) are, for instance, labeled EVT in the corpus whereas this lexeme appears as an OBJECT in our lexicon, while this lexeme denotes an activity and thus an event. Other mistakes can be observed, such as: ADMINISTRATION#2 (in the sense of *set of persons in charge of the administration of something*) which has been described as occurrences of ADMINISTRATION#1 (*the resulting state of the process*). Confronting data extracted from our corpus and data from our lexicon thus allows us to ensure their quality.

6. Conclusion

We have presented in this paper a corpus-based semantic annotation project. The resulting annotated corpus is the groundwork for one of the main outcomes of the project: a semantic and syntactic electronic lexicon for French deverbal nouns, linked to their occurrences in the French Treebank[9]. This lexicon will be the first, so far as we know, to propose a description of aspectual properties for French nouns, in the continuity of projects such as Nomlex (Macleod et al., 1998) and SIMPLE (Bel et al., 2000).

By combining theoretical and empirical approaches to linguistic description, the Nomage project provides stable data available for further analysis regarding nominal aspect. The interaction between both approaches has proven its interest. On the one hand, theory provides the empirical approach with linguistic tests and an ontology. On the other hand, the theoretical approach is challenged by contextual data, which raise the question of vagueness and of the relevance of the theoretical classes.

The relationship between the verbal and nominal aspectual systems also has to be further investigated. There are structural differences, due to the grammatical specificities of each category, that should be questioned. For instance, as long as there are no OBJECT verbs, under which conditions do verbs yield OBJECT nominalizations? Does the mass-count nominal feature correspond to some lexical property in the verbal domain? How can the cases of conversion (e.g. MARCHE MARCHER) be analyzed with regard to aspectual inheritance?

In future work, we intend to extend the semantic annotation process to French deadjectival nouns (e.g. FIDÉLITÉ from FIDÈLE), and to non deverbal predicative nouns (e.g. crime). We also intend to extend our methodology to other languages: Spanish, English and Catalan.

7. References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks, Building and Using Parsed Corpora*. Kluwer, Dordrecht.

Lucie Barque, Richard Huyghe, Anne Jugnet, and Rafael Marín. 2009. Two types of deverbal activity nouns in French. In *5th International Conference on Generative Approaches to the Lexicon*, pages 169–175, Pisa.

Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowsky, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. In *Proceedings of LREC 2000*, pages 1379–1384, Athens.

Antonietta Bisetto and Chiara Melloni. 2008. On the Interpretation of Nominals: Towards a Result-Oriented Verb Classification. In *Proceedings of the 40th Linguistics Colloquium*, Frankfurt.

Noam Chomsky. 1970. Remarks on nominalizations. In A.J. Roderick & P.S. Rosenbaum, editor, *Readings in English Tranformational Grammar*. Ginn and Co, Waltham (MA).

Georgette Dal and Fiammetta Namer. 2010. Les noms en *-ance/-ence* du français : quel(s) patron(s) constructionnel(s)? In *Actes en ligne du 2e Congrès Mondial de Linguistique Française*, pages 893–907, La Nouvelle Orléans, États-Unis.

David Dowty. 1979. *Word Meaning and Montague Grammar*. D. Reidel Publishing Co., Dordrecht.

Nelly Flaux and Danièle Van de Velde. 2000. *Les noms en français : esquisse de classement*. Ophrys, Paris.

Livio Gaeta. 2002. *Quando i verbi compaiono come nomi. Un saggio di morfologia naturale*. FrancoAngeli, Milano.

Danièle Godard and Jacques Jayez. 1996. Types nominaux et anaphores : le cas des objets et des événements. In W. De Mulder, L. Tasmowki-De Ryck, and C. Vetters, editors, *Cahiers Chronos 1*.

Jane Grimshaw. 1990. *Argument structure*. MIT Press, Cambridge, MA.

Pauline Haas and Richard Huyghe. 2010. Les propriétés aspectuelles des noms d'activités. *Cahiers Chronos*, 21.

Pauline Haas, Richard Huyghe, and Rafael Marín. 2008. Du verbe au nom : calques et décalages aspectuels. In *Congrès Mondial de Linguistique Française (CMLF 2008)*, pages 2039–2053, Paris.

Sophie Heyd and Marie-Laurence Knittel. 2009. Les noms d'activité parmi les noms abstraits : propriétés aspectuelles, distributionnelles et interprétatives. *Lingvisticæ investigationes*, 32-1.

Robert B. Lees. 1960. *The Grammar of English Nominalizations*. Mouton, The Hague.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barret, and Ruth Reeves. 1998. NOMLEX: A Lexicon of Nominalizations. In *Proceedings of Euralex'98*, Liege, Belgium.

Igor Mel'čuk. 2004a. Actants in Semantics and Syntax I: actants in semantics. *Linguistics*, 42(1):1–66.

Igor Mel'čuk. 2004b. Actants in Semantics and Syntax II: actants in syntax. *Linguistics*, 42(2):247–291.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekeley, Veronkia Zielinska, and Brian Young. 2004. The Cross-Breeding of Dictionaries. In *Proceedings of LREC-2004*, Lisbon, Portugal.

Adam Meyers. 2007. Annotation Guidelines for NomBank. *Online publication:*

[9] A simple query interface to the Nomage lexicon can be accessed at `http://nomage.recherche.univ-lille3.fr/webgui`.

http://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf.

Jasmina Milicévic and Alain Polguère. 2010. Ambivalence sémantique des noms de communication langagière en français. In *Congrès Mondial de Linguistique Française (CMLF 2010)*, Paris.

Aina Peris, Mariona Taulé, and Horacio Rodríguez. 2010. Semantic Annotation of Deverbal Nominalizations in the Spanish corpus AnCora. In *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, University of Tartu, Estonia.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

Ludovic Tanguy and Nabil Hathout. 2002. Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In *Actes de TALN 2002*, Nancy.

M. Taulé, M.A. Martí, and M. Recasens. 2008. Ancora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of 6th International Conference on Language Resources and Evaluation*, Marrakesh, Morocco.

Karel Van den Eynde and Piet Mertens. 2003. La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104.